Online Learning
Lecture 3

Nicolò Cesa-Bianchi

Università degli Studi di Milano

FTRL recap

FTRL:

$$\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{V}} \psi(\boldsymbol{w}) + \sum_{s=1}^{t} \ell_s(\boldsymbol{w}_s)$$

Best model in $\mathbb{V}$

$$\boldsymbol{w}^* = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{V}} \sum_{t=1}^{T} \ell_t(\boldsymbol{w})$$

Regret

$$R_T = \sum_{t=1}^{T} \ell_t(\boldsymbol{w}_t) - \sum_{t=1}^{T} \ell_t(\boldsymbol{w}^*)$$

FTRL = FTL + regularization

- $L_t = \ell_0 + \ell_1 + \cdots + \ell_t$ where $\ell_0 = \psi$

FTRL = FTL + regularization

- $L_t = \ell_0 + \ell_1 + \cdots + \ell_t$ where $\ell_0 = \psi$
- $\boldsymbol{w}_1 = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{V}} \ell_0(\boldsymbol{w}) = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{V}} \psi(\boldsymbol{w})$

FTRL = FTL + regularization

- $L_t = \ell_0 + \ell_1 + \cdots + \ell_t$ where $\ell_0 = \psi$
- $\boldsymbol{w}_1 = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{V}} \ell_0(\boldsymbol{w}) = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{V}} \psi(\boldsymbol{w})$
- $\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{V}} L_t(\boldsymbol{w})$

FTRL = FTL + regularization

- $L_t = \ell_0 + \ell_1 + \cdots + \ell_t$ where $\ell_0 = \psi$
- $\boldsymbol{w}_1 = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{V}} \ell_0(\boldsymbol{w}) = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{V}} \psi(\boldsymbol{w})$
- $\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{V}} L_t(\boldsymbol{w})$
- We give a different proof of the FTL stability lemma

FTRL stability lemma

$$\sum_{t=0}^{T} \ell_t(\boldsymbol{w}_{t+1}) \leq \inf_{\boldsymbol{u} \in \mathbb{V}} \sum_{t=0}^{T} \ell_t(\boldsymbol{u})$$

(we prove this by induction on T)

FTRL stability lemma

$$\sum_{t=0}^{T} \ell_t(\boldsymbol{w}_{t+1}) \le \inf_{\boldsymbol{u}\in\mathbb{V}} \sum_{t=0}^{T} \ell_t(\boldsymbol{u})$$

(we prove this by induction on T)

$$\ell_0(\boldsymbol{w}_1) \le \inf_{\boldsymbol{u}\in\mathbb{V}} \ell_0(\boldsymbol{u})$$

(base case $T = 0$: $\boldsymbol{w}_1 = \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{V}} \ell_0(\boldsymbol{w})$)

FTRL stability lemma

$$\sum_{t=0}^{T} \ell_t(\boldsymbol{w}_{t+1}) \le \inf_{\boldsymbol{u}\in\mathbb{V}} \sum_{t=0}^{T} \ell_t(\boldsymbol{u})$$

(we prove this by induction on T)

$$\ell_0(\boldsymbol{w}_1) \le \inf_{\boldsymbol{u}\in\mathbb{V}} \ell_0(\boldsymbol{u})$$

(base case $T = 0$: $\boldsymbol{w}_1 = \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{V}} \ell_0(\boldsymbol{w})$)

$$\sum_{t=0}^{T-1} \ell_t(\boldsymbol{w}_{t+1}) \le \inf_{\boldsymbol{u}\in\mathbb{V}} \sum_{t=0}^{T-1} \ell_t(\boldsymbol{u})$$

$(T-1 \to T)$

FTRL stability lemma

$$\sum_{t=0}^{T} \ell_t(\boldsymbol{w}_{t+1}) \le \inf_{\boldsymbol{u}\in\mathbb{V}} \sum_{t=0}^{T} \ell_t(\boldsymbol{u})$$ (we prove this by induction on T)

$$\ell_0(\boldsymbol{w}_1) \le \inf_{\boldsymbol{u}\in\mathbb{V}} \ell_0(\boldsymbol{u})$$ (base case $T = 0$: $\boldsymbol{w}_1 = \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{V}} \ell_0(\boldsymbol{w})$)

$$\sum_{t=0}^{T-1} \ell_t(\boldsymbol{w}_{t+1}) \le \inf_{\boldsymbol{u}\in\mathbb{V}} \sum_{t=0}^{T-1} \ell_t(\boldsymbol{u})$$ ($T - 1 \to T$)

$$\sum_{t=0}^{T-1} \ell_t(\boldsymbol{w}_{t+1}) \le \sum_{t=0}^{T-1} \ell_t(\boldsymbol{w}_{T+1})$$ (choose $\boldsymbol{u} = \boldsymbol{w}_{T+1}$)

FTRL stability lemma

$$\sum_{t=0}^{T} \ell_t(\boldsymbol{w}_{t+1}) \le \inf_{\boldsymbol{u}\in\mathbb{V}} \sum_{t=0}^{T} \ell_t(\boldsymbol{u})$$ (we prove this by induction on T)

$$\ell_0(\boldsymbol{w}_1) \le \inf_{\boldsymbol{u}\in\mathbb{V}} \ell_0(\boldsymbol{u})$$ (base case $T = 0$: $\boldsymbol{w}_1 = \operatorname{argmin}_{\boldsymbol{w}\in\mathbb{V}} \ell_0(\boldsymbol{w})$)

$$\sum_{t=0}^{T-1} \ell_t(\boldsymbol{w}_{t+1}) \le \inf_{\boldsymbol{u}\in\mathbb{V}} \sum_{t=0}^{T-1} \ell_t(\boldsymbol{u})$$ ($T - 1 \to T$)

$$\sum_{t=0}^{T-1} \ell_t(\boldsymbol{w}_{t+1}) \le \sum_{t=0}^{T-1} \ell_t(\boldsymbol{w}_{T+1})$$ (choose $\boldsymbol{u} = \boldsymbol{w}_{T+1}$)

$$\sum_{t=0}^{T} \ell_t(\boldsymbol{w}_{t+1}) \le \sum_{t=0}^{T} \ell_t(\boldsymbol{w}_{T+1}) = \inf_{\boldsymbol{u}\in\mathbb{V}} \sum_{t=0}^{T} \ell_t(\boldsymbol{u})$$ (add $\ell_T(\boldsymbol{w}_{T+1})$ on both sides)

FTRL stability lemma

$$\sum_{t=0}^{T} \ell_t(\boldsymbol{w}_{t+1}) \le \inf_{\boldsymbol{u}\in\mathbb{V}} \sum_{t=0}^{T} \ell_t(\boldsymbol{u}) \qquad \text{(we prove this by induction on } T\text{)}$$

$$\ell_0(\boldsymbol{w}_1) \le \inf_{\boldsymbol{u}\in\mathbb{V}} \ell_0(\boldsymbol{u}) \qquad \text{(base case } T=0\text{: } \boldsymbol{w}_1 = \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{V}} \ell_0(\boldsymbol{w})\text{)}$$

$$\sum_{t=0}^{T-1} \ell_t(\boldsymbol{w}_{t+1}) \le \inf_{\boldsymbol{u}\in\mathbb{V}} \sum_{t=0}^{T-1} \ell_t(\boldsymbol{u}) \qquad (T-1 \to T)$$

$$\sum_{t=0}^{T-1} \ell_t(\boldsymbol{w}_{t+1}) \le \sum_{t=0}^{T-1} \ell_t(\boldsymbol{w}_{T+1}) \qquad \text{(choose } \boldsymbol{u} = \boldsymbol{w}_{T+1}\text{)}$$

$$\sum_{t=0}^{T} \ell_t(\boldsymbol{w}_{t+1}) \le \sum_{t=0}^{T} \ell_t(\boldsymbol{w}_{T+1}) = \inf_{\boldsymbol{u}\in\mathbb{V}} \sum_{t=0}^{T} \ell_t(\boldsymbol{u}) \qquad \text{(add } \ell_T(\boldsymbol{w}_{T+1}) \text{ on both sides)}$$

$$R_T = \sum_{t=1}^{T} \Big(\ell_t(\boldsymbol{w}_t) - \ell_t(\boldsymbol{w}^*)\Big) \le \ell_0(\boldsymbol{w}^*) - \ell_0(\boldsymbol{w}_1) + \sum_{t=1}^{T} \Big(\ell_t(\boldsymbol{w}_t) - \ell_t(\boldsymbol{w}_{t+1})\Big)$$

FTRL stability analysis

- Assume ψ is μ-strongly convex with respect to $\|\cdot\|$

FTRL stability analysis

- Assume ψ is μ-strongly convex with respect to $\|\cdot\|$
- Pick a learning rate $\eta > 0$ and consider $\frac{\psi}{\eta}$ which is μ/η-strongly convex

FTRL stability analysis

- Assume ψ is μ-strongly convex with respect to $\|\cdot\|$
- Pick a learning rate $\eta > 0$ and consider $\frac{\psi}{\eta}$ which is μ/η-strongly convex
- For all $t \geq 1$, we assume ℓ_t is G-Lipschitz with respect to $\|\cdot\|$

FTRL stability analysis

- Assume ψ is μ-strongly convex with respect to $\|\cdot\|$
- Pick a learning rate $\eta > 0$ and consider $\frac{\psi}{\eta}$ which is μ/η-strongly convex
- For all $t \geq 1$, we assume ℓ_t is G-Lipschitz with respect to $\|\cdot\|$
- $\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{V}} L_t(\boldsymbol{w})$

FTRL stability analysis

- Assume ψ is μ-strongly convex with respect to $\|\cdot\|$
- Pick a learning rate $\eta > 0$ and consider $\frac{\psi}{\eta}$ which is μ/η-strongly convex
- For all $t \geq 1$, we assume ℓ_t is G-Lipschitz with respect to $\|\cdot\|$
- $\boldsymbol{w}_{t+1} = \underset{\boldsymbol{w} \in \mathbb{V}}{\operatorname{argmin}} \, L_t(\boldsymbol{w})$

$$L_{t-1}(\boldsymbol{w}_{t+1}) - L_{t-1}(\boldsymbol{w}_t) \geq \nabla L_{t-1}(\boldsymbol{w}_t)^\top (\boldsymbol{w}_{t+1} - \boldsymbol{w}_t) + \frac{\mu}{2\eta} \|\boldsymbol{w}_t - \boldsymbol{w}_{t+1}\|^2 \geq \frac{\mu}{2\eta} \|\boldsymbol{w}_t - \boldsymbol{w}_{t+1}\|^2$$

FTRL stability analysis

- Assume ψ is μ-strongly convex with respect to $\|\cdot\|$
- Pick a learning rate $\eta > 0$ and consider $\frac{\psi}{\eta}$ which is μ/η-strongly convex
- For all $t \geq 1$, we assume ℓ_t is G-Lipschitz with respect to $\|\cdot\|$
- $\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{V}} L_t(\boldsymbol{w})$

$$L_{t-1}(\boldsymbol{w}_{t+1}) - L_{t-1}(\boldsymbol{w}_t) \geq \nabla L_{t-1}(\boldsymbol{w}_t)^\top (\boldsymbol{w}_{t+1} - \boldsymbol{w}_t) + \frac{\mu}{2\eta} \|\boldsymbol{w}_t - \boldsymbol{w}_{t+1}\|^2 \geq \frac{\mu}{2\eta} \|\boldsymbol{w}_t - \boldsymbol{w}_{t+1}\|^2$$

$$L_t(\boldsymbol{w}_t) - L_t(\boldsymbol{w}_{t+1}) \geq \nabla L_t(\boldsymbol{w}_{t+1})^\top (\boldsymbol{w}_t - \boldsymbol{w}_{t+1}) + \frac{\mu}{2\eta} \|\boldsymbol{w}_t - \boldsymbol{w}_{t+1}\|^2 \geq \frac{\mu}{2\eta} \|\boldsymbol{w}_t - \boldsymbol{w}_{t+1}\|^2$$

FTRL stability analysis

- Assume ψ is μ-strongly convex with respect to $\|\cdot\|$
- Pick a learning rate $\eta > 0$ and consider $\frac{\psi}{\eta}$ which is μ/η-strongly convex
- For all $t \geq 1$, we assume ℓ_t is G-Lipschitz with respect to $\|\cdot\|$
- $\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{V}} L_t(\boldsymbol{w})$

$$L_{t-1}(\boldsymbol{w}_{t+1}) - L_{t-1}(\boldsymbol{w}_t) \geq \nabla L_{t-1}(\boldsymbol{w}_t)^\top (\boldsymbol{w}_{t+1} - \boldsymbol{w}_t) + \frac{\mu}{2\eta} \|\boldsymbol{w}_t - \boldsymbol{w}_{t+1}\|^2 \geq \frac{\mu}{2\eta} \|\boldsymbol{w}_t - \boldsymbol{w}_{t+1}\|^2$$

$$L_t(\boldsymbol{w}_t) - L_t(\boldsymbol{w}_{t+1}) \geq \nabla L_t(\boldsymbol{w}_{t+1})^\top (\boldsymbol{w}_t - \boldsymbol{w}_{t+1}) + \frac{\mu}{2\eta} \|\boldsymbol{w}_t - \boldsymbol{w}_{t+1}\|^2 \geq \frac{\mu}{2\eta} \|\boldsymbol{w}_t - \boldsymbol{w}_{t+1}\|^2$$

$$\ell_t(\boldsymbol{w}_t) - \ell_t(\boldsymbol{w}_{t+1}) \geq \frac{\mu}{\eta} \|\boldsymbol{w}_t - \boldsymbol{w}_{t+1}\|^2 \qquad \text{(by summing the two above inequalities)}$$

FTRL stability analysis

- Assume ψ is μ-strongly convex with respect to $\|\cdot\|$
- Pick a learning rate $\eta > 0$ and consider $\frac{\psi}{\eta}$ which is μ/η-strongly convex
- For all $t \geq 1$, we assume ℓ_t is G-Lipschitz with respect to $\|\cdot\|$
- $\boldsymbol{w}_{t+1} = \underset{\boldsymbol{w} \in \mathbb{V}}{\operatorname{argmin}}\, L_t(\boldsymbol{w})$

$$L_{t-1}(\boldsymbol{w}_{t+1}) - L_{t-1}(\boldsymbol{w}_t) \geq \nabla L_{t-1}(\boldsymbol{w}_t)^\top(\boldsymbol{w}_{t+1} - \boldsymbol{w}_t) + \frac{\mu}{2\eta}\|\boldsymbol{w}_t - \boldsymbol{w}_{t+1}\|^2 \geq \frac{\mu}{2\eta}\|\boldsymbol{w}_t - \boldsymbol{w}_{t+1}\|^2$$

$$L_t(\boldsymbol{w}_t) - L_t(\boldsymbol{w}_{t+1}) \geq \nabla L_t(\boldsymbol{w}_{t+1})^\top(\boldsymbol{w}_t - \boldsymbol{w}_{t+1}) + \frac{\mu}{2\eta}\|\boldsymbol{w}_t - \boldsymbol{w}_{t+1}\|^2 \geq \frac{\mu}{2\eta}\|\boldsymbol{w}_t - \boldsymbol{w}_{t+1}\|^2$$

$$\ell_t(\boldsymbol{w}_t) - \ell_t(\boldsymbol{w}_{t+1}) \geq \frac{\mu}{\eta}\|\boldsymbol{w}_t - \boldsymbol{w}_{t+1}\|^2 \qquad \text{(by summing the two above inequalities)}$$

$$\ell_t(\boldsymbol{w}_t) - \ell_t(\boldsymbol{w}_{t+1}) \leq G\|\boldsymbol{w}_t - \boldsymbol{w}_{t+1}\| \qquad (G\text{-Lipschitzness of } \ell_t \text{ for } t \geq 1)$$

FTRL stability analysis

- Assume ψ is μ-strongly convex with respect to $\|\cdot\|$
- Pick a learning rate $\eta > 0$ and consider $\frac{\psi}{\eta}$ which is μ/η-strongly convex
- For all $t \geq 1$, we assume ℓ_t is G-Lipschitz with respect to $\|\cdot\|$
- $\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{V}} L_t(\boldsymbol{w})$

$$L_{t-1}(\boldsymbol{w}_{t+1}) - L_{t-1}(\boldsymbol{w}_t) \geq \nabla L_{t-1}(\boldsymbol{w}_t)^\top (\boldsymbol{w}_{t+1} - \boldsymbol{w}_t) + \frac{\mu}{2\eta} \|\boldsymbol{w}_t - \boldsymbol{w}_{t+1}\|^2 \geq \frac{\mu}{2\eta} \|\boldsymbol{w}_t - \boldsymbol{w}_{t+1}\|^2$$

$$L_t(\boldsymbol{w}_t) - L_t(\boldsymbol{w}_{t+1}) \geq \nabla L_t(\boldsymbol{w}_{t+1})^\top (\boldsymbol{w}_t - \boldsymbol{w}_{t+1}) + \frac{\mu}{2\eta} \|\boldsymbol{w}_t - \boldsymbol{w}_{t+1}\|^2 \geq \frac{\mu}{2\eta} \|\boldsymbol{w}_t - \boldsymbol{w}_{t+1}\|^2$$

$$\ell_t(\boldsymbol{w}_t) - \ell_t(\boldsymbol{w}_{t+1}) \geq \frac{\mu}{\eta} \|\boldsymbol{w}_t - \boldsymbol{w}_{t+1}\|^2 \qquad \text{(by summing the two above inequalities)}$$

$$\ell_t(\boldsymbol{w}_t) - \ell_t(\boldsymbol{w}_{t+1}) \leq G \|\boldsymbol{w}_t - \boldsymbol{w}_{t+1}\| \qquad (G\text{-Lipschitzness of } \ell_t \text{ for } t \geq 1)$$

$$\|\boldsymbol{w}_t - \boldsymbol{w}_{t+1}\| \leq \frac{G\eta}{\mu} \quad \text{implying} \quad \ell_t(\boldsymbol{w}_t) - \ell_t(\boldsymbol{w}_{t+1}) \leq \frac{G^2\eta}{\mu}$$

FTRL regret bound

$$R_T \le \frac{\psi(\boldsymbol{w}^*) - \psi(\boldsymbol{w}_1)}{\eta} + \sum_{t=1}^{T} \Big(\ell_t(\boldsymbol{w}_t) - \ell_t(\boldsymbol{w}_{t+1}) \Big) \le \frac{\psi(\boldsymbol{w}^*) - \psi(\boldsymbol{w}_1)}{\eta} + \eta \frac{G^2}{\mu} T$$

FTRL regret bound

$$R_T \le \frac{\psi(\boldsymbol{w}^*) - \psi(\boldsymbol{w}_1)}{\eta} + \sum_{t=1}^{T} \Big(\ell_t(\boldsymbol{w}_t) - \ell_t(\boldsymbol{w}_{t+1}) \Big) \le \frac{\psi(\boldsymbol{w}^*) - \psi(\boldsymbol{w}_1)}{\eta} + \eta \frac{G^2}{\mu} T$$

Assuming $\max_{\boldsymbol{u} \in \mathbb{V}} \psi(\boldsymbol{u}) - \min_{\boldsymbol{w} \in \mathbb{V}} \psi(\boldsymbol{w}) = D^2$ and choosing $\eta = \frac{D}{G}\sqrt{\frac{\mu}{T}}$ we get

$$R_T \le 2GD\sqrt{\frac{T}{\mu}} = \mathcal{O}(GD\sqrt{T})$$

Lipschitz constants and dual norm of gradients

- Dual norm of $\|\cdot\|$ is $\|\boldsymbol{w}\|_* = \sup_{\boldsymbol{u}\in\mathbb{R}^d} \frac{\boldsymbol{w}^\top \boldsymbol{u}}{\|\boldsymbol{u}\|}$

Lipschitz constants and dual norm of gradients

- Dual norm of $\|\cdot\|$ is $\|\boldsymbol{w}\|_* = \sup_{\boldsymbol{u}\in\mathbb{R}^d} \frac{\boldsymbol{w}^\top \boldsymbol{u}}{\|\boldsymbol{u}\|}$
- Example (p-norms): the dual norm of $\|\cdot\|_p$ is $\|\cdot\|_q$ where $\frac{1}{p} + \frac{1}{q} = 1$ and $p, q > 1$

Lipschitz constants and dual norm of gradients

- Dual norm of $\|\cdot\|$ is $\|\boldsymbol{w}\|_* = \sup_{\boldsymbol{u}\in\mathbb{R}^d} \dfrac{\boldsymbol{w}^\top \boldsymbol{u}}{\|\boldsymbol{u}\|}$
- Example (p-norms): the dual norm of $\|\cdot\|_p$ is $\|\cdot\|_q$ where $\frac{1}{p}+\frac{1}{q}=1$ and $p,q>1$
- Relationship to convex duality: $\psi = \frac{1}{2}\|\cdot\|_p^2$ and $\psi^* = \frac{1}{2}\|\cdot\|_q^2$

Lipschitz constants and dual norm of gradients

- Dual norm of $\|\cdot\|$ is $\|\boldsymbol{w}\|_* = \sup_{\boldsymbol{u} \in \mathbb{R}^d} \dfrac{\boldsymbol{w}^\top \boldsymbol{u}}{\|\boldsymbol{u}\|}$
- Example (p-norms): the dual norm of $\|\cdot\|_p$ is $\|\cdot\|_q$ where $\frac{1}{p} + \frac{1}{q} = 1$ and $p, q > 1$
- Relationship to convex duality: $\psi = \frac{1}{2}\|\cdot\|_p^2$ and $\psi^* = \frac{1}{2}\|\cdot\|_q^2$
- Hölder inequality: $\boldsymbol{u}^\top \boldsymbol{w} \le \|\boldsymbol{u}\| \, \|\boldsymbol{w}\|_*$
- Fenchel-Young inequality: $\boldsymbol{u}^\top \boldsymbol{w} \le \psi(\boldsymbol{u}) + \psi^*(\boldsymbol{w})$

Lipschitz constants and dual norm of gradients

- Dual norm of $\|\cdot\|$ is $\|\boldsymbol{w}\|_* = \sup_{\boldsymbol{u}\in\mathbb{R}^d} \dfrac{\boldsymbol{w}^\top \boldsymbol{u}}{\|\boldsymbol{u}\|}$
- Example (p-norms): the dual norm of $\|\cdot\|_p$ is $\|\cdot\|_q$ where $\frac{1}{p} + \frac{1}{q} = 1$ and $p, q > 1$
- Relationship to convex duality: $\psi = \frac{1}{2}\|\cdot\|_p^2$ and $\psi^* = \frac{1}{2}\|\cdot\|_q^2$
- Hölder inequality: $\boldsymbol{u}^\top \boldsymbol{w} \le \|\boldsymbol{u}\| \, \|\boldsymbol{w}\|_*$
- Fenchel-Young inequality: $\boldsymbol{u}^\top \boldsymbol{w} \le \psi(\boldsymbol{u}) + \psi^*(\boldsymbol{w})$

Theorem

Let $\ell : \mathbb{V} \to \mathbb{R}$ be a differentiable convex function. Then ℓ is G-Lipschitz over $\mathbb{V}$ with respect to a norm $\|\cdot\|$ iff for all $\boldsymbol{w} \in \mathbb{V}$ we have that $\|\nabla \ell(\boldsymbol{w})\|_* \le G$

Lipschitz constants and dual norm of gradients

- Dual norm of $\|\cdot\|$ is $\|\boldsymbol{w}\|_* = \sup_{\boldsymbol{u}\in\mathbb{R}^d} \frac{\boldsymbol{w}^\top \boldsymbol{u}}{\|\boldsymbol{u}\|}$
- Example (p-norms): the dual norm of $\|\cdot\|_p$ is $\|\cdot\|_q$ where $\frac{1}{p}+\frac{1}{q}=1$ and $p,q>1$
- Relationship to convex duality: $\psi = \frac{1}{2}\|\cdot\|_p^2$ and $\psi^* = \frac{1}{2}\|\cdot\|_q^2$
- Hölder inequality: $\boldsymbol{u}^\top \boldsymbol{w} \le \|\boldsymbol{u}\|\,\|\boldsymbol{w}\|_*$
- Fenchel-Young inequality: $\boldsymbol{u}^\top \boldsymbol{w} \le \psi(\boldsymbol{u}) + \psi^*(\boldsymbol{w})$

Theorem

Let $\ell : \mathbb{V} \to \mathbb{R}$ be a differentiable convex function. Then ℓ is G-Lipschitz over $\mathbb{V}$ with respect to a norm $\|\cdot\|$ iff for all $\boldsymbol{w} \in \mathbb{V}$ we have that $\|\nabla\ell(\boldsymbol{w})\|_* \le G$

- Assume $\max_i \left|(\nabla\ell)_i\right| = \Theta(1)$

Lipschitz constants and dual norm of gradients

- Dual norm of $\|\cdot\|$ is $\|\boldsymbol{w}\|_* = \sup_{\boldsymbol{u}\in\mathbb{R}^d} \dfrac{\boldsymbol{w}^\top \boldsymbol{u}}{\|\boldsymbol{u}\|}$
- Example (p-norms): the dual norm of $\|\cdot\|_p$ is $\|\cdot\|_q$ where $\frac{1}{p}+\frac{1}{q}=1$ and $p,q>1$
- Relationship to convex duality: $\psi = \frac{1}{2}\|\cdot\|_p^2$ and $\psi^* = \frac{1}{2}\|\cdot\|_q^2$
- Hölder inequality: $\boldsymbol{u}^\top \boldsymbol{w} \le \|\boldsymbol{u}\|\,\|\boldsymbol{w}\|_*$
- Fenchel-Young inequality: $\boldsymbol{u}^\top \boldsymbol{w} \le \psi(\boldsymbol{u}) + \psi^*(\boldsymbol{w})$

Theorem

Let $\ell : \mathbb{V} \to \mathbb{R}$ be a differentiable convex function. Then ℓ is G-Lipschitz over $\mathbb{V}$ with respect to a norm $\|\cdot\|$ iff for all $\boldsymbol{w} \in \mathbb{V}$ we have that $\|\nabla \ell(\boldsymbol{w})\|_* \le G$

- Assume $\max_i |(\nabla \ell)_i| = \Theta(1)$
- If ℓ is G-Lipschitz with respect to $\|\cdot\|_2$, then $G = \mathcal{O}(\sqrt{d})$

Lipschitz constants and dual norm of gradients

- Dual norm of $\|\cdot\|$ is $\|\boldsymbol{w}\|_* = \sup_{\boldsymbol{u}\in\mathbb{R}^d} \frac{\boldsymbol{w}^\top \boldsymbol{u}}{\|\boldsymbol{u}\|}$
- Example (p-norms): the dual norm of $\|\cdot\|_p$ is $\|\cdot\|_q$ where $\frac{1}{p}+\frac{1}{q}=1$ and $p,q>1$
- Relationship to convex duality: $\psi = \frac{1}{2}\|\cdot\|_p^2$ and $\psi^* = \frac{1}{2}\|\cdot\|_q^2$
- Hölder inequality: $\boldsymbol{u}^\top \boldsymbol{w} \le \|\boldsymbol{u}\|\,\|\boldsymbol{w}\|_*$
- Fenchel-Young inequality: $\boldsymbol{u}^\top \boldsymbol{w} \le \psi(\boldsymbol{u}) + \psi^*(\boldsymbol{w})$

Theorem

Let $\ell : \mathbb{V} \to \mathbb{R}$ be a differentiable convex function. Then ℓ is G-Lipschitz over $\mathbb{V}$ with respect to a norm $\|\cdot\|$ iff for all $\boldsymbol{w} \in \mathbb{V}$ we have that $\|\nabla\ell(\boldsymbol{w})\|_* \le G$

- Assume $\max_i |(\nabla\ell)_i| = \Theta(1)$
- If ℓ is G-Lipschitz with respect to $\|\cdot\|_2$, then $G = \mathcal{O}(\sqrt{d})$
- If ℓ is G-Lipschitz with respect to $\|\cdot\|_1$, then $G = \mathcal{O}(1)$

Matching the regularizer to the geometry of the model space

Projected Lazy OGD

Matching the regularizer to the geometry of the model space

Projected Lazy OGD

- Take $\mathbb{V}$ to be the closed Euclidean ball of radius D

Matching the regularizer to the geometry of the model space

Projected Lazy OGD

- Take $\mathbb{V}$ to be the closed Euclidean ball of radius D
- $\psi = \frac{1}{2}\|\cdot\|_2^2$ is 1-strongly convex with respect to $\|\cdot\|_2$

Matching the regularizer to the geometry of the model space

Projected Lazy OGD

- Take $\mathbb{V}$ to be the closed Euclidean ball of radius D
- $\psi = \frac{1}{2}\|\cdot\|_2^2$ is 1-strongly convex with respect to $\|\cdot\|_2$
- $\max_{\boldsymbol{u}\in\mathbb{V}} \psi(\boldsymbol{u}) - \min_{\boldsymbol{w}\in\mathbb{V}} \psi(\boldsymbol{w}) = \frac{D^2}{2}$

Matching the regularizer to the geometry of the model space

Projected Lazy OGD

- Take $\mathbb{V}$ to be the closed Euclidean ball of radius D
- $\psi = \dfrac{1}{2}\|\cdot\|_2^2$ is 1-strongly convex with respect to $\|\cdot\|_2$
- $\displaystyle\max_{\boldsymbol{u}\in\mathbb{V}} \psi(\boldsymbol{u}) - \min_{\boldsymbol{w}\in\mathbb{V}} \psi(\boldsymbol{w}) = \frac{D^2}{2}$
- Lipschitz constant: $G = \mathcal{O}(\sqrt{d})$

Matching the regularizer to the geometry of the model space

Projected Lazy OGD

- Take $\mathbb{V}$ to be the closed Euclidean ball of radius D
- $\psi = \frac{1}{2}\|\cdot\|_2^2$ is 1-strongly convex with respect to $\|\cdot\|_2$
- $\max_{\boldsymbol{u}\in\mathbb{V}} \psi(\boldsymbol{u}) - \min_{\boldsymbol{w}\in\mathbb{V}} \psi(\boldsymbol{w}) = \frac{D^2}{2}$
- Lipschitz constant: $G = \mathcal{O}(\sqrt{d})$
- $R_T \le 2GD\sqrt{\frac{T}{\mu}} = \mathcal{O}(D\sqrt{dT})$

Matching the regularizer to the geometry of the model space

EG

Matching the regularizer to the geometry of the model space

EG

- $\mathbb{V}$ is probability simplex Δ_d

Matching the regularizer to the geometry of the model space

EG

- $\mathbb{V}$ is probability simplex Δ_d
- $\psi(\boldsymbol{p}) = \sum_i p_i \ln p_i$ is 1-strongly convex with respect to $\|\cdot\|_1$

Matching the regularizer to the geometry of the model space

EG

- $\mathbb{V}$ is probability simplex Δ_d
- $\psi(\boldsymbol{p}) = \sum_i p_i \ln p_i$ is 1-strongly convex with respect to $\|\cdot\|_1$
- $D = \max_{\boldsymbol{u}\in\mathbb{V}} \psi(\boldsymbol{u}) - \min_{\boldsymbol{w}\in\mathbb{V}} \psi(\boldsymbol{w}) = \ln d$

Matching the regularizer to the geometry of the model space

EG

- $\mathbb{V}$ is probability simplex Δ_d
- $\psi(\boldsymbol{p}) = \sum_i p_i \ln p_i$ is 1-strongly convex with respect to $\|\cdot\|_1$
- $D = \max_{\boldsymbol{u}\in\mathbb{V}} \psi(\boldsymbol{u}) - \min_{\boldsymbol{w}\in\mathbb{V}} \psi(\boldsymbol{w}) = \ln d$
- Lipschitz constant: $G = \mathcal{O}(1)$

Matching the regularizer to the geometry of the model space

EG

- $\mathbb{V}$ is probability simplex Δ_d
- $\psi(\boldsymbol{p}) = \sum_i p_i \ln p_i$ is 1-strongly convex with respect to $\|\cdot\|_1$
- $D = \max_{\boldsymbol{u}\in\mathbb{V}} \psi(\boldsymbol{u}) - \min_{\boldsymbol{w}\in\mathbb{V}} \psi(\boldsymbol{w}) = \ln d$
- Lipschitz constant: $G = \mathcal{O}(1)$
- $R_T \le 2GD\sqrt{\frac{T}{\mu}} = \mathcal{O}\big(\sqrt{T \ln d}\big)$

Matching the regularizer to the geometry of the model space

EG

- $\mathbb{V}$ is probability simplex Δ_d
- $\psi(\boldsymbol{p}) = \sum_i p_i \ln p_i$ is 1-strongly convex with respect to $\|\cdot\|_1$
- $D = \max_{\boldsymbol{u}\in\mathbb{V}} \psi(\boldsymbol{u}) - \min_{\boldsymbol{w}\in\mathbb{V}} \psi(\boldsymbol{w}) = \ln d$
- Lipschitz constant: $G = \mathcal{O}(1)$
- $R_T \le 2GD\sqrt{\frac{T}{\mu}} = \mathcal{O}(\sqrt{T \ln d})$
- For $\mathbb{V} = \Delta_d$, projected lazy OGD only achieves $R_T = \mathcal{O}(\sqrt{dT})$
- The geometry of $\mathbb{V}$ matters

Lower bounds

- $\mathbb{V}$ is a bounded set of Euclidean diameter D

Lower bounds

- $\mathbb{V}$ is a bounded set of Euclidean diameter D
- Take $\boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathbb{V}$ such that $\|\boldsymbol{v}_1 - \boldsymbol{v}_2\|_2 = D$ and set $\boldsymbol{z}_0 = (\boldsymbol{v}_1 - \boldsymbol{v}_2) / \|\boldsymbol{v}_1 - \boldsymbol{v}_2\|_2$

Lower bounds

- $\mathbb{V}$ is a bounded set of Euclidean diameter D
- Take $\boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathbb{V}$ such that $\|\boldsymbol{v}_1 - \boldsymbol{v}_2\|_2 = D$ and set $\boldsymbol{z}_0 = (\boldsymbol{v}_1 - \boldsymbol{v}_2) / \|\boldsymbol{v}_1 - \boldsymbol{v}_2\|_2$
- G-Lipschitz linear losses: $\ell_t(\boldsymbol{w}) = \varepsilon_t G\, \boldsymbol{w}^\top \boldsymbol{z}_0$ where $\varepsilon_t \in \{-1, 1\}$ are uniform i.i.d.

Lower bounds

- $\mathbb{V}$ is a bounded set of Euclidean diameter D
- Take $\boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathbb{V}$ such that $\|\boldsymbol{v}_1 - \boldsymbol{v}_2\|_2 = D$ and set $\boldsymbol{z}_0 = (\boldsymbol{v}_1 - \boldsymbol{v}_2) / \|\boldsymbol{v}_1 - \boldsymbol{v}_2\|_2$
- G-Lipschitz linear losses: $\ell_t(\boldsymbol{w}) = \varepsilon_t G\, \boldsymbol{w}^\top \boldsymbol{z}_0$ where $\varepsilon_t \in \{-1, 1\}$ are uniform i.i.d.

$$\mathbb{E}\left[\max_{\boldsymbol{u} \in \{\boldsymbol{v}_1, \boldsymbol{v}_2\}} R_T(\boldsymbol{u})\right]$$

Lower bounds

- $\mathbb{V}$ is a bounded set of Euclidean diameter D
- Take $\boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathbb{V}$ such that $\|\boldsymbol{v}_1 - \boldsymbol{v}_2\|_2 = D$ and set $\boldsymbol{z}_0 = (\boldsymbol{v}_1 - \boldsymbol{v}_2)/\|\boldsymbol{v}_1 - \boldsymbol{v}_2\|_2$
- G-Lipschitz linear losses: $\ell_t(\boldsymbol{w}) = \varepsilon_t G\, \boldsymbol{w}^\top \boldsymbol{z}_0$ where $\varepsilon_t \in \{-1, 1\}$ are uniform i.i.d.

$$\mathbb{E}\left[\max_{\boldsymbol{u}\in\{\boldsymbol{v}_1,\boldsymbol{v}_2\}} R_T(\boldsymbol{u})\right] = \mathbb{E}\left[\max_{\boldsymbol{u}\in\{\boldsymbol{v}_1,\boldsymbol{v}_2\}} \sum_{t=1}^T \ell_t(\boldsymbol{u})\right] \qquad \text{(since } \mathbb{E}[\ell_t(\boldsymbol{w})] = 0\text{)}$$

Lower bounds

- $\mathbb{V}$ is a bounded set of Euclidean diameter D
- Take $\boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathbb{V}$ such that $\|\boldsymbol{v}_1 - \boldsymbol{v}_2\|_2 = D$ and set $\boldsymbol{z}_0 = (\boldsymbol{v}_1 - \boldsymbol{v}_2) / \|\boldsymbol{v}_1 - \boldsymbol{v}_2\|_2$
- G-Lipschitz linear losses: $\ell_t(\boldsymbol{w}) = \varepsilon_t G\, \boldsymbol{w}^\top \boldsymbol{z}_0$ where $\varepsilon_t \in \{-1, 1\}$ are uniform i.i.d.

$$\mathbb{E}\left[\max_{\boldsymbol{u}\in\{\boldsymbol{v}_1,\boldsymbol{v}_2\}} R_T(\boldsymbol{u})\right] = \mathbb{E}\left[\max_{\boldsymbol{u}\in\{\boldsymbol{v}_1,\boldsymbol{v}_2\}} \sum_{t=1}^T \ell_t(\boldsymbol{u})\right] \qquad \text{(since } \mathbb{E}[\ell_t(\boldsymbol{w})] = 0\text{)}$$

$$= \frac{G}{2}\mathbb{E}\left[\left|\sum_{t=1}^T \varepsilon_t \boldsymbol{z}_0^\top (\boldsymbol{v}_1 - \boldsymbol{v}_2)\right|\right] \qquad \text{(using } \max\{a,b\} = \tfrac{1}{2}(a+b+|a-b|)\text{)}$$

Lower bounds

- $\mathbb{V}$ is a bounded set of Euclidean diameter D
- Take $\boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathbb{V}$ such that $\|\boldsymbol{v}_1 - \boldsymbol{v}_2\|_2 = D$ and set $\boldsymbol{z}_0 = (\boldsymbol{v}_1 - \boldsymbol{v}_2)/\|\boldsymbol{v}_1 - \boldsymbol{v}_2\|_2$
- G-Lipschitz linear losses: $\ell_t(\boldsymbol{w}) = \varepsilon_t G\, \boldsymbol{w}^\top \boldsymbol{z}_0$ where $\varepsilon_t \in \{-1, 1\}$ are uniform i.i.d.

$$\mathbb{E}\left[\max_{\boldsymbol{u}\in\{\boldsymbol{v}_1,\boldsymbol{v}_2\}} R_T(\boldsymbol{u})\right] = \mathbb{E}\left[\max_{\boldsymbol{u}\in\{\boldsymbol{v}_1,\boldsymbol{v}_2\}} \sum_{t=1}^{T} \ell_t(\boldsymbol{u})\right] \qquad \text{(since } \mathbb{E}[\ell_t(\boldsymbol{w})] = 0\text{)}$$

$$= \frac{G}{2}\mathbb{E}\left[\left|\sum_{t=1}^{T} \varepsilon_t \boldsymbol{z}_0^\top (\boldsymbol{v}_1 - \boldsymbol{v}_2)\right|\right] \qquad \text{(using } \max\{a,b\} = \tfrac{1}{2}(a+b+|a-b|)\text{)}$$

$$= \frac{GD}{2}\mathbb{E}\left[\left|\sum_{t=1}^{T} \varepsilon_t\right|\right] \qquad \text{(because } \boldsymbol{z}_0^\top(\boldsymbol{v}_1 - \boldsymbol{v}_2) = D\text{)}$$

Lower bounds

- $\mathbb{V}$ is a bounded set of Euclidean diameter D
- Take $\boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathbb{V}$ such that $\|\boldsymbol{v}_1 - \boldsymbol{v}_2\|_2 = D$ and set $\boldsymbol{z}_0 = (\boldsymbol{v}_1 - \boldsymbol{v}_2)/\|\boldsymbol{v}_1 - \boldsymbol{v}_2\|_2$
- G-Lipschitz linear losses: $\ell_t(\boldsymbol{w}) = \varepsilon_t G\, \boldsymbol{w}^\top \boldsymbol{z}_0$ where $\varepsilon_t \in \{-1, 1\}$ are uniform i.i.d.

$$\mathbb{E}\left[\max_{\boldsymbol{u}\in\{\boldsymbol{v}_1,\boldsymbol{v}_2\}} R_T(\boldsymbol{u})\right] = \mathbb{E}\left[\max_{\boldsymbol{u}\in\{\boldsymbol{v}_1,\boldsymbol{v}_2\}} \sum_{t=1}^{T} \ell_t(\boldsymbol{u})\right] \qquad \text{(since } \mathbb{E}[\ell_t(\boldsymbol{w})] = 0\text{)}$$

$$= \frac{G}{2}\mathbb{E}\left[\left|\sum_{t=1}^{T} \varepsilon_t \boldsymbol{z}_0^\top (\boldsymbol{v}_1 - \boldsymbol{v}_2)\right|\right] \qquad \text{(using } \max\{a,b\} = \tfrac{1}{2}(a+b+|a-b|)\text{)}$$

$$= \frac{GD}{2}\mathbb{E}\left[\left|\sum_{t=1}^{T} \varepsilon_t\right|\right] \qquad \text{(because } \boldsymbol{z}_0^\top (\boldsymbol{v}_1 - \boldsymbol{v}_2) = D\text{)}$$

$$\geq GD\sqrt{\frac{T}{8}} \qquad \text{(Khintchine inequality)}$$

Lower bound for the simplex

- Stochastic linear losses $\boldsymbol{\ell}_t = (\ell_t(1), \dots, \ell_t(d))$

Lower bound for the simplex

- Stochastic linear losses $\boldsymbol{\ell}_t = (\ell_t(1), \ldots, \ell_t(d))$
- $\ell_t(i) \in \{0, 1\}$ independent random coin flip for all $t \geq 1$ and $i = 1, \ldots, d$

Lower bound for the simplex

- Stochastic linear losses $\boldsymbol{\ell}_t = (\ell_t(1), \dots, \ell_t(d))$
- $\ell_t(i) \in \{0,1\}$ independent random coin flip for all $t \ge 1$ and $i = 1, \dots, d$
- For any online algorithm $\mathbb{E}\left[\sum_{t=1}^{T} \boldsymbol{p}_t^\top \boldsymbol{\ell}_t\right] = \frac{T}{2}$

Lower bound for the simplex

- Stochastic linear losses $\boldsymbol{\ell}_t = (\ell_t(1), \dots, \ell_t(d))$
- $\ell_t(i) \in \{0,1\}$ independent random coin flip for all $t \ge 1$ and $i = 1, \dots, d$
- For any online algorithm $\mathbb{E}\left[\sum_{t=1}^{T} \boldsymbol{p}_t^\top \boldsymbol{\ell}_t\right] = \frac{T}{2}$
- Then the expected regret is

$$\frac{T}{2} - \mathbb{E}\left[\min_{\boldsymbol{p}\in\Delta_d} \sum_{t=1}^{T} \boldsymbol{q}^\top \boldsymbol{\ell}_t\right] = \frac{T}{2} - \mathbb{E}\left[\min_{i=1,\dots,d} \sum_{t=1}^{T} \ell_t(i)\right]$$

$$= \mathbb{E}\left[\max_{i=1,\dots,d} \sum_{t=1}^{T} \left(\frac{1}{2} - \ell_t(i)\right)\right]$$

$$= \frac{1}{2}\mathbb{E}\left[\max_{i=1,\dots,d} \sum_{t=1}^{T} \varepsilon_t(i)\right]$$

where $\varepsilon_t(i) \in \{-1, 1\}$ are uniform i.i.d.

Lower bound for the simplex: finishing up

$$\mathbb{E}\left[\max_{i=1,\dots,d}\sum_{t=1}^{T}\varepsilon_t(i)\right] = \left(1 - o(1)\right)\sqrt{2T\ln d}$$

Therefore, $R_T = \Omega\left(\sqrt{T\ln d}\right)$

FTRL with time-varying regularizer

- Fix a sequence $\psi_1, \psi_2, \dots$ of regularizers

FTRL with time-varying regularizer

- Fix a sequence $\psi_1, \psi_2, \dots$ of regularizers
- $L_t^\psi = \psi_{t+1} + L_t = \psi_{t+1} + \ell_1 + \cdots + \ell_t$

FTRL with time-varying regularizer

- Fix a sequence $\psi_1, \psi_2, \ldots$ of regularizers
- $L_t^{\psi} = \psi_{t+1} + L_t = \psi_{t+1} + \ell_1 + \cdots + \ell_t$
- FTRL prediction $\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{V}} L_t^{\psi}(\boldsymbol{w})$ and best model $\boldsymbol{w}^* = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{V}} \sum_{t=1}^{T} \ell_t(\boldsymbol{w})$

FTRL with time-varying regularizer

- Fix a sequence $\psi_1, \psi_2, \dots$ of regularizers
- $L_t^\psi = \psi_{t+1} + L_t = \psi_{t+1} + \ell_1 + \dots + \ell_t$
- FTRL prediction $\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{V}} L_t^\psi(\boldsymbol{w})$ and best model $\boldsymbol{w}^* = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{V}} \sum_{t=1}^{T} \ell_t(\boldsymbol{w})$

$$-L_T(\boldsymbol{w}^*) = \psi_{T+1}(\boldsymbol{w}^*) - L_T^\psi(\boldsymbol{w}^*)$$

FTRL with time-varying regularizer

- Fix a sequence $\psi_1, \psi_2, \dots$ of regularizers
- $L_t^\psi = \psi_{t+1} + L_t = \psi_{t+1} + \ell_1 + \cdots + \ell_t$
- FTRL prediction $\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{V}} L_t^\psi(\boldsymbol{w})$ and best model $\boldsymbol{w}^* = \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{V}} \sum_{t=1}^T \ell_t(\boldsymbol{w})$

$$-L_T(\boldsymbol{w}^*) = \psi_{T+1}(\boldsymbol{w}^*) - L_T^\psi(\boldsymbol{w}^*)$$

$$-L_T(\boldsymbol{w}^*) = \psi_{T+1}(\boldsymbol{w}^*) - \underbrace{L_0^\psi(\boldsymbol{w}_1)}_{=\psi_1(\boldsymbol{w}_1)} + \underbrace{L_0^\psi(\boldsymbol{w}_1) - L_T^\psi(\boldsymbol{w}_{T+1})}_{\text{write as telescoping}} + \underbrace{L_T^\psi(\boldsymbol{w}_{T+1}) - L_T^\psi(\boldsymbol{w}^*)}_{\le 0}$$

FTRL with time-varying regularizer

- Fix a sequence $\psi_1, \psi_2, \dots$ of regularizers
- $L_t^\psi = \psi_{t+1} + L_t = \psi_{t+1} + \ell_1 + \cdots + \ell_t$
- FTRL prediction $\boldsymbol{w}_{t+1} = \underset{\boldsymbol{w}\in\mathbb{V}}{\operatorname{argmin}}\, L_t^\psi(\boldsymbol{w})$ and best model $\boldsymbol{w}^* = \underset{\boldsymbol{w}\in\mathbb{V}}{\operatorname{argmin}} \sum_{t=1}^{T} \ell_t(\boldsymbol{w})$

$$-L_T(\boldsymbol{w}^*) = \psi_{T+1}(\boldsymbol{w}^*) - L_T^\psi(\boldsymbol{w}^*)$$

$$-L_T(\boldsymbol{w}^*) = \psi_{T+1}(\boldsymbol{w}^*) - \underbrace{L_0^\psi(\boldsymbol{w}_1)}_{=\psi_1(\boldsymbol{w}_1)} + \underbrace{L_0^\psi(\boldsymbol{w}_1) - L_T^\psi(\boldsymbol{w}_{T+1})}_{\text{write as telescoping}} + \underbrace{L_T^\psi(\boldsymbol{w}_{T+1}) - L_T^\psi(\boldsymbol{w}^*)}_{\le 0}$$

$$-L_T(\boldsymbol{w}^*) \le \psi_{T+1}(\boldsymbol{w}^*) - \psi_1(\boldsymbol{w}_1) + \sum_{t=1}^{T} \Big(L_{t-1}^\psi(\boldsymbol{w}_t) - L_t^\psi(\boldsymbol{w}_{t+1}) \Big)$$

FTRL with time-varying regularizer

- Fix a sequence $\psi_1, \psi_2, \dots$ of regularizers
- $L_t^\psi = \psi_{t+1} + L_t = \psi_{t+1} + \ell_1 + \cdots + \ell_t$
- FTRL prediction $\boldsymbol{w}_{t+1} = \underset{\boldsymbol{w} \in \mathbb{V}}{\operatorname{argmin}}\, L_t^\psi(\boldsymbol{w})$ and best model $\boldsymbol{w}^* = \underset{\boldsymbol{w} \in \mathbb{V}}{\operatorname{argmin}} \sum_{t=1}^{T} \ell_t(\boldsymbol{w})$

$$-L_T(\boldsymbol{w}^*) = \psi_{T+1}(\boldsymbol{w}^*) - L_T^\psi(\boldsymbol{w}^*)$$

$$-L_T(\boldsymbol{w}^*) = \psi_{T+1}(\boldsymbol{w}^*) - \underbrace{L_0^\psi(\boldsymbol{w}_1)}_{=\psi_1(\boldsymbol{w}_1)} + \underbrace{L_0^\psi(\boldsymbol{w}_1) - L_T^\psi(\boldsymbol{w}_{T+1})}_{\text{write as telescoping}} + \underbrace{L_T^\psi(\boldsymbol{w}_{T+1}) - L_T^\psi(\boldsymbol{w}^*)}_{\le 0}$$

$$-L_T(\boldsymbol{w}^*) \le \psi_{T+1}(\boldsymbol{w}^*) - \psi_1(\boldsymbol{w}_1) + \sum_{t=1}^{T} \Big(L_{t-1}^\psi(\boldsymbol{w}_t) - L_t^\psi(\boldsymbol{w}_{t+1}) \Big)$$

$$R_T \le \psi_{T+1}(\boldsymbol{w}^*) - \psi_1(\boldsymbol{w}_1) + \sum_{t=1}^{T} \Big(L_{t-1}^\psi(\boldsymbol{w}_t) - L_t^\psi(\boldsymbol{w}_{t+1}) + \ell_t(\boldsymbol{w}_t) \Big)$$

Analysis of FTRL with time-varying regularizer

$$R_T \leq \psi_{T+1}(\boldsymbol{w}^*) - \psi_1(\boldsymbol{w}_1) + \sum_{t=1}^{T} \Big(L_{t-1}^{\psi}(\boldsymbol{w}_t) - L_t^{\psi}(\boldsymbol{w}_{t+1}) + \ell_t(\boldsymbol{w}_t) \Big)$$

Analysis of FTRL with time-varying regularizer

$$R_T \le \psi_T(\boldsymbol{w}^*) - \psi_1(\boldsymbol{w}_1) + \sum_{t=1}^{T} \Big(L_{t-1}^{\psi}(\boldsymbol{w}_t) - L_t^{\psi}(\boldsymbol{w}_{t+1}) + \ell_t(\boldsymbol{w}_t) \Big)$$

Analysis of FTRL with time-varying regularizer

$$R_T \le \psi_T(\boldsymbol{w}^*) - \psi_1(\boldsymbol{w}_1) + \sum_{t=1}^{T} \Big(L_{t-1}^{\psi}(\boldsymbol{w}_t) - L_t^{\psi}(\boldsymbol{w}_{t+1}) + \ell_t(\boldsymbol{w}_t) \Big)$$

We now bound the terms $L_{t-1}^{\psi}(\boldsymbol{w}_t) - L_t^{\psi}(\boldsymbol{w}_{t+1}) + \ell_t(\boldsymbol{w}_t)$

Analysis of FTRL with time-varying regularizer (cont.)

- Assume ψ_t is μ_t-strongly convex and $\psi_t \le \psi_{t+1}$ for $t \ge 1$

Analysis of FTRL with time-varying regularizer (cont.)

- Assume ψ_t is μ_t-strongly convex and $\psi_t \le \psi_{t+1}$ for $t \ge 1$
- Recall $f(\boldsymbol{w}) - f(\boldsymbol{w}^*) \ge \frac{\mu}{2}\|\boldsymbol{w} - \boldsymbol{w}^*\|^2$ for f μ-strongly convex and $\boldsymbol{w}^* = \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{V}} f(\boldsymbol{w})$

Analysis of FTRL with time-varying regularizer (cont.)

- Assume ψ_t is μ_t-strongly convex and $\psi_t \leq \psi_{t+1}$ for $t \geq 1$
- Recall $f(\boldsymbol{w}) - f(\boldsymbol{w}^*) \geq \dfrac{\mu}{2}\|\boldsymbol{w} - \boldsymbol{w}^*\|^2$ for f μ-strongly convex and $\boldsymbol{w}^* = \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{V}} f(\boldsymbol{w})$
- Let $\boldsymbol{w}_t^* = \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{V}} L_{t-1}^{\psi}(\boldsymbol{w}) + \ell_t(\boldsymbol{w})$

Analysis of FTRL with time-varying regularizer (cont.)

- Assume ψ_t is μ_t-strongly convex and $\psi_t \le \psi_{t+1}$ for $t \ge 1$
- Recall $f(\boldsymbol{w}) - f(\boldsymbol{w}^*) \ge \frac{\mu}{2} \|\boldsymbol{w} - \boldsymbol{w}^*\|^2$ for f μ-strongly convex and $\boldsymbol{w}^* = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{V}} f(\boldsymbol{w})$
- Let $\boldsymbol{w}_t^* = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{V}} L_{t-1}^{\psi}(\boldsymbol{w}) + \ell_t(\boldsymbol{w})$

$$
\begin{aligned}
& L_{t-1}^{\psi}(\boldsymbol{w}_t) - L_{t}^{\psi}(\boldsymbol{w}_{t+1}) + \ell_t(\boldsymbol{w}_t) \\
& \quad = \left(L_{t-1}^{\psi}(\boldsymbol{w}_t) + \ell_t(\boldsymbol{w}_t)\right) - \left(L_{t-1}^{\psi}(\boldsymbol{w}_{t+1}) + \ell_t(\boldsymbol{w}_{t+1})\right) + \psi_t(\boldsymbol{w}_{t+1}) - \psi_{t+1}(\boldsymbol{w}_{t+1})
\end{aligned}
$$

Analysis of FTRL with time-varying regularizer (cont.)

- Assume ψ_t is μ_t-strongly convex and $\psi_t \le \psi_{t+1}$ for $t \ge 1$
- Recall $f(\boldsymbol{w}) - f(\boldsymbol{w}^*) \ge \frac{\mu}{2} \|\boldsymbol{w} - \boldsymbol{w}^*\|^2$ for f μ-strongly convex and $\boldsymbol{w}^* = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{V}} f(\boldsymbol{w})$
- Let $\boldsymbol{w}_t^* = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{V}} L_{t-1}^{\psi}(\boldsymbol{w}) + \ell_t(\boldsymbol{w})$

$$
\begin{aligned}
& L_{t-1}^{\psi}(\boldsymbol{w}_t) - L_{t}^{\psi}(\boldsymbol{w}_{t+1}) + \ell_t(\boldsymbol{w}_t) \\
& \quad = \left(L_{t-1}^{\psi}(\boldsymbol{w}_t) + \ell_t(\boldsymbol{w}_t)\right) - \left(L_{t-1}^{\psi}(\boldsymbol{w}_{t+1}) + \ell_t(\boldsymbol{w}_{t+1})\right) + \psi_t(\boldsymbol{w}_{t+1}) - \psi_{t+1}(\boldsymbol{w}_{t+1}) \\
& \quad \le \left(L_{t-1}^{\psi}(\boldsymbol{w}_t) + \ell_t(\boldsymbol{w}_t)\right) - \left(L_{t-1}^{\psi}(\boldsymbol{w}_t^*) + \ell_t(\boldsymbol{w}_t^*)\right) \quad \text{(minimality of } \boldsymbol{w}_t^* \text{ and condition on } \psi_t\text{)}
\end{aligned}
$$

Analysis of FTRL with time-varying regularizer (cont.)

- Assume ψ_t is μ_t-strongly convex and $\psi_t \le \psi_{t+1}$ for $t \ge 1$
- Recall $f(\boldsymbol{w}) - f(\boldsymbol{w}^*) \ge \frac{\mu}{2}\|\boldsymbol{w} - \boldsymbol{w}^*\|^2$ for f μ-strongly convex and $\boldsymbol{w}^* = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{V}} f(\boldsymbol{w})$
- Let $\boldsymbol{w}_t^* = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{V}} L_{t-1}^{\psi}(\boldsymbol{w}) + \ell_t(\boldsymbol{w})$

$$
\begin{aligned}
& L_{t-1}^{\psi}(\boldsymbol{w}_t) - L_{t}^{\psi}(\boldsymbol{w}_{t+1}) + \ell_t(\boldsymbol{w}_t) \\
& \quad = \left(L_{t-1}^{\psi}(\boldsymbol{w}_t) + \ell_t(\boldsymbol{w}_t)\right) - \left(L_{t-1}^{\psi}(\boldsymbol{w}_{t+1}) + \ell_t(\boldsymbol{w}_{t+1})\right) + \psi_t(\boldsymbol{w}_{t+1}) - \psi_{t+1}(\boldsymbol{w}_{t+1}) \\
& \quad \le \left(L_{t-1}^{\psi}(\boldsymbol{w}_t) + \ell_t(\boldsymbol{w}_t)\right) - \left(L_{t-1}^{\psi}(\boldsymbol{w}_t^*) + \ell_t(\boldsymbol{w}_t^*)\right) \quad \text{(minimality of } \boldsymbol{w}_t^* \text{ and condition on } \psi_t\text{)}
\end{aligned}
$$

$$
\left(L_{t-1}^{\psi}(\boldsymbol{w}_t) + \ell_t(\boldsymbol{w}_t)\right) - \left(L_{t-1}^{\psi}(\boldsymbol{w}_t^*) + \ell_t(\boldsymbol{w}_t^*)\right) \ge \frac{\mu_t}{2}\|\boldsymbol{w}_t - \boldsymbol{w}_t^*\|^2 \quad \text{(s.c. + minimality of } \boldsymbol{w}_t^*\text{)}
$$

Analysis of FTRL with time-varying regularizer (cont.)

- Assume ψ_t is μ_t-strongly convex and $\psi_t \le \psi_{t+1}$ for $t \ge 1$
- Recall $f(\boldsymbol{w}) - f(\boldsymbol{w}^*) \ge \frac{\mu}{2}\|\boldsymbol{w} - \boldsymbol{w}^*\|^2$ for f μ-strongly convex and $\boldsymbol{w}^* = \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{V}} f(\boldsymbol{w})$
- Let $\boldsymbol{w}_t^* = \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{V}} L_{t-1}^{\psi}(\boldsymbol{w}) + \ell_t(\boldsymbol{w})$

$$
\begin{aligned}
& L_{t-1}^{\psi}(\boldsymbol{w}_t) - L_{t}^{\psi}(\boldsymbol{w}_{t+1}) + \ell_t(\boldsymbol{w}_t) \\
&\quad = \big(L_{t-1}^{\psi}(\boldsymbol{w}_t) + \ell_t(\boldsymbol{w}_t)\big) - \big(L_{t-1}^{\psi}(\boldsymbol{w}_{t+1}) + \ell_t(\boldsymbol{w}_{t+1})\big) + \psi_t(\boldsymbol{w}_{t+1}) - \psi_{t+1}(\boldsymbol{w}_{t+1}) \\
&\quad \le \big(L_{t-1}^{\psi}(\boldsymbol{w}_t) + \ell_t(\boldsymbol{w}_t)\big) - \big(L_{t-1}^{\psi}(\boldsymbol{w}_t^*) + \ell_t(\boldsymbol{w}_t^*)\big) \quad \text{(minimality of } \boldsymbol{w}_t^* \text{ and condition on } \psi_t\text{)}
\end{aligned}
$$

$$
\big(L_{t-1}^{\psi}(\boldsymbol{w}_t) + \ell_t(\boldsymbol{w}_t)\big) - \big(L_{t-1}^{\psi}(\boldsymbol{w}_t^*) + \ell_t(\boldsymbol{w}_t^*)\big) \ge \frac{\mu_t}{2}\|\boldsymbol{w}_t - \boldsymbol{w}_t^*\|^2 \quad \text{(s.c. + minimality of } \boldsymbol{w}_t^*\text{)}
$$

$$
L_{t-1}^{\psi}(\boldsymbol{w}_t) - L_{t-1}^{\psi}(\boldsymbol{w}_t^*) + \ell_t(\boldsymbol{w}_t) - \ell_t(\boldsymbol{w}_t^*) \le G\|\boldsymbol{w}_t - \boldsymbol{w}_t^*\| \quad \text{(minimality of } \boldsymbol{w}_t \text{ + Lip.)}
$$

Analysis of FTRL with time-varying regularizer (cont.)

- Assume ψ_t is μ_t-strongly convex and $\psi_t \leq \psi_{t+1}$ for $t \geq 1$
- Recall $f(\boldsymbol{w}) - f(\boldsymbol{w}^*) \geq \frac{\mu}{2} \|\boldsymbol{w} - \boldsymbol{w}^*\|^2$ for f μ-strongly convex and $\boldsymbol{w}^* = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{V}} f(\boldsymbol{w})$
- Let $\boldsymbol{w}_t^* = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{V}} L_{t-1}^{\psi}(\boldsymbol{w}) + \ell_t(\boldsymbol{w})$

$$
\begin{aligned}
& L_{t-1}^{\psi}(\boldsymbol{w}_t) - L_{t}^{\psi}(\boldsymbol{w}_{t+1}) + \ell_t(\boldsymbol{w}_t) \\
&\quad = \left(L_{t-1}^{\psi}(\boldsymbol{w}_t) + \ell_t(\boldsymbol{w}_t)\right) - \left(L_{t-1}^{\psi}(\boldsymbol{w}_{t+1}) + \ell_t(\boldsymbol{w}_{t+1})\right) + \psi_t(\boldsymbol{w}_{t+1}) - \psi_{t+1}(\boldsymbol{w}_{t+1}) \\
&\quad \leq \left(L_{t-1}^{\psi}(\boldsymbol{w}_t) + \ell_t(\boldsymbol{w}_t)\right) - \left(L_{t-1}^{\psi}(\boldsymbol{w}_t^*) + \ell_t(\boldsymbol{w}_t^*)\right) \quad \text{(minimality of } \boldsymbol{w}_t^* \text{ and condition on } \psi_t)
\end{aligned}
$$

$$
\left(L_{t-1}^{\psi}(\boldsymbol{w}_t) + \ell_t(\boldsymbol{w}_t)\right) - \left(L_{t-1}^{\psi}(\boldsymbol{w}_t^*) + \ell_t(\boldsymbol{w}_t^*)\right) \geq \frac{\mu_t}{2} \|\boldsymbol{w}_t - \boldsymbol{w}_t^*\|^2 \quad \text{(s.c. + minimality of } \boldsymbol{w}_t^*)
$$

$$
L_{t-1}^{\psi}(\boldsymbol{w}_t) - L_{t-1}^{\psi}(\boldsymbol{w}_t^*) + \ell_t(\boldsymbol{w}_t) - \ell_t(\boldsymbol{w}_t^*) \leq G \|\boldsymbol{w}_t - \boldsymbol{w}_t^*\| \quad \text{(minimality of } \boldsymbol{w}_t + \text{Lip.)}
$$

$$
L_{t-1}^{\psi}(\boldsymbol{w}_t) - L_{t}^{\psi}(\boldsymbol{w}_{t+1}) + \ell_t(\boldsymbol{w}_t) \leq \frac{2G^2}{\mu_t}
$$

Regret bound

Assume $\psi \geq 0$ is μ-strongly convex and $\psi_t = \dfrac{\psi}{\eta_{t-1}}$ where $\eta_t \leq \eta_{t-1}$ for $t \geq 1$

Regret bound

Assume $\psi \geq 0$ is μ-strongly convex and $\psi_t = \dfrac{\psi}{\eta_{t-1}}$ where $\eta_t \leq \eta_{t-1}$ for $t \geq 1$

$$R_T \leq \psi_T(\boldsymbol{w}^*) + 2G^2 \sum_{t=1}^{T} \frac{1}{\mu_t}$$

Regret bound

Assume $\psi \geq 0$ is μ-strongly convex and $\psi_t = \dfrac{\psi}{\eta_{t-1}}$ where $\eta_t \leq \eta_{t-1}$ for $t \geq 1$

$$
\begin{aligned}
R_T &\leq \psi_T(\boldsymbol{w}^*) + 2G^2 \sum_{t=1}^{T} \frac{1}{\mu_t} \\
&= \frac{D^2}{\eta_T} + 2G^2 \sum_{t=1}^{T} \frac{\eta_{t-1}}{\mu}
\end{aligned}
$$

(ψ_t is (μ/η_{t-1})-strongly convex)

Regret bound

Assume $\psi \geq 0$ is μ-strongly convex and $\psi_t = \dfrac{\psi}{\eta_{t-1}}$ where $\eta_t \leq \eta_{t-1}$ for $t \geq 1$

$$R_T \leq \psi_T(\boldsymbol{w}^*) + 2G^2 \sum_{t=1}^{T} \frac{1}{\mu_t}$$

$$= \frac{D^2}{\eta_T} + 2G^2 \sum_{t=1}^{T} \frac{\eta_{t-1}}{\mu} \qquad (\psi_t \text{ is } (\mu/\eta_{t-1})\text{-strongly convex})$$

$$= GD\sqrt{\frac{T}{\mu}} + 2GD\sqrt{\frac{1}{\mu}} \sum_{t=1}^{T} \sqrt{t} \qquad \left(\eta_{t-1} = \frac{D}{G}\sqrt{\frac{\mu}{t}}\right)$$

Regret bound

Assume $\psi \geq 0$ is μ-strongly convex and $\psi_t = \dfrac{\psi}{\eta_{t-1}}$ where $\eta_t \leq \eta_{t-1}$ for $t \geq 1$

$$
\begin{aligned}
R_T &\leq \psi_T(\boldsymbol{w}^*) + 2G^2 \sum_{t=1}^{T} \frac{1}{\mu_t} \\
&= \frac{D^2}{\eta_T} + 2G^2 \sum_{t=1}^{T} \frac{\eta_{t-1}}{\mu} && (\psi_t \text{ is } (\mu/\eta_{t-1})\text{-strongly convex}) \\
&= GD\sqrt{\frac{T}{\mu}} + 2GD\sqrt{\frac{1}{\mu}} \sum_{t=1}^{T} \sqrt{t} && \left(\eta_{t-1} = \frac{D}{G}\sqrt{\frac{\mu}{t}}\right) \\
&\leq 5GD\sqrt{\frac{T}{\mu}}
\end{aligned}
$$

using $\displaystyle\sum_{t=1}^{T} \sqrt{t} \leq 2\sqrt{T}$